Movement similarity assessment using symbolic representation of trajectories

Dodge, S ; Laube, P ; Weibel, Robert

Abstract: This paper describes a novel approach for finding similar trajectories, using trajectory segmentation based on movement parameters such as speed, acceleration, or direction. First, a segmentation technique is applied to decompose trajectories into a set of segments with homogeneous characteristics with respect to a particular movement parameter. Each segment is assigned to a movement parameter class, representing the behavior of the movement parameter. Accordingly, the segmentation procedure transforms a trajectory to a sequence of class labels, that is, a symbolic representation. A modified version of edit distance, called Normalized Weighted Edit Distance (NWED) is introduced as a similarity measure between different sequences. As an application, we demonstrate how the method can be employed to cluster trajectories. The performance of the approach is assessed in two case studies using real movement datasets from two different application domains, namely, North Atlantic Hurricane trajectories and GPS tracks of couriers in London. Three different experiments have been conducted that respond to different facets of the proposed techniques, and that compare our NWED measure to a related method.

DOI: https://doi.org/10.1080/13658816.2011.630003

International Journal of Geographical Information Science, 2011, in press, pre-print version

RESEARCH ARTICLE

Movement Similarity Assessment Using Symbolic Representation of Trajectories

Somayeh Dodge*, Patrick Laube, and Robert Weibel

University of Zurich, Winterthurerstrasse 190, 8057, Zurich, Switzerland.

This paper describes a novel approach for finding similar trajectories, using trajectory segmentation based on movement parameters such as speed, acceleration, or direction. First, a segmentation technique is applied to decompose trajectories into a set of segments with homogeneous characteristics with respect to a particular movement parameter. Each segment is assigned to a movement parameter class, representing the behavior of the movement parameter. Accordingly, the segmentation procedure transforms a trajectory to a sequence of class labels, that is, a symbolic representation. A modified version of edit distance, called Normalized Weighted Edit Distance (NWED) is introduced as a similarity measure between different sequences. As an application, we demonstrate how the method can be employed to cluster trajectories. The performance of the approach is assessed in two case studies using real movement datasets from two different application domains, namely, North Atlantic Hurricane trajectories and GPS tracks of couriers in London. Three different experiments have been conducted that respond to different facets of the proposed techniques, and that compare our NWED measure to a related method.

Keywords: Movement similarity; trajectory segmentation; movement parameter; movement patterns; trajectory clustering.

* Corresponding author. Email: somayeh.dodge@geo.uzh.ch

1. Introduction

Understanding the dynamic behavior of moving objects (e.g., humans, animals, vehicles, etc.) or processes (e.g., hurricanes, oil spills) is quickly becoming a key task in many GIScience application domains. In movement behavior studies, it is essential to take into account the key parameters that characterize the movement of objects, so-called *movement parameters (MP)* such as speed, acceleration, or direction (Dodge *et al.* 2008). For example, in intelligent transport systems it is important to know the speed patterns of vehicles across the street network at different times of the day in order to detect traffic anomalies and incidences (Miller and Han 2009). Likewise, in order to predict the position or time of a hurricane's landfall it is crucial to know the speed, acceleration and direction behavior of hurricanes before and close to the landfall (Elsner and Kara 1999). Movement parameters either can be derived from the trajectories of objects or recorded directly by sensors. Today, with the emergence of new sensor technologies such as accelerometers, gyroscopes, and recent advances of in-vehicle sensors, a variety of movement parameters of mobile objects can be registered as an object moves. The development of analysis techniques that are capable of exploiting these new sources of information thus appears to be a logical step forward for knowledge discovery from movement datasets.

Similarity analysis as an exploratory tool in movement research is an important and challenging topic. Recently, similarity analysis has become the focus of many studies in mobility data mining (Giannotti and Pedreschi 2008, Miller and Han 2009). A review of the relevant literature suggests that although there are several trajectory similarity search methods that are relatively well developed, most of them are restricted to geometric abstractions of the objects' movement path as a static curve (i.e. a time-ordered sequence of coordinates). And only a few of the available similarity analysis techniques take the variations of movement parameters into account. However, in many applications spatial similarity alone may not be appropriate to detect objects with similar movement characteristics. For instance, trajectories of vehicles that move on the same route are similar in terms of geometric shape (i.e. spatial similarity). However, the speed variations of vehicles might exhibit different patterns over time that cannot be discovered with purely spatial similarity assessment. As another example, "history has shown that many of the hurricanes that have struck New England over the last 100 years share very similar characteristics"[1]. For instance, in 1954 two hurricanes, Carol and Edna, made landfall on Cape Cod, MA. and exhibited very similar movement characteristics insofar as they have been known as identical hurricanes in the meteorological literature in terms of their evolution (Malkin and Holzworth 1954). However, since their movement paths exhibit a distinctively different geometry (see Figure 1), these two hurricanes would not be extracted as similar using the available spatial movement similarity measures. By the same token, hurricane Edna and hurricane Isaac, that their movement paths follow a similar geometry as seen in Figure 1, would be extracted as similar, although their speeds exhibit a very different pattern. Therefore, in addition to geometric similarity of trajectories, it seems inevitable to investigate the similarity of the *variation of movement parameters* over time.

The *objective* of this paper is to contribute to trajectory similarity search, by proposing a new approach that allows seeking for trajectories in movement datasets that exhibit *common patterns in the variation of their movement parameters over time*. The *main contribution* of this research is the introduction of a novel technique for spatio-temporal

[1] http://www.hurricanescience.org/history/storms/1950s/carol/

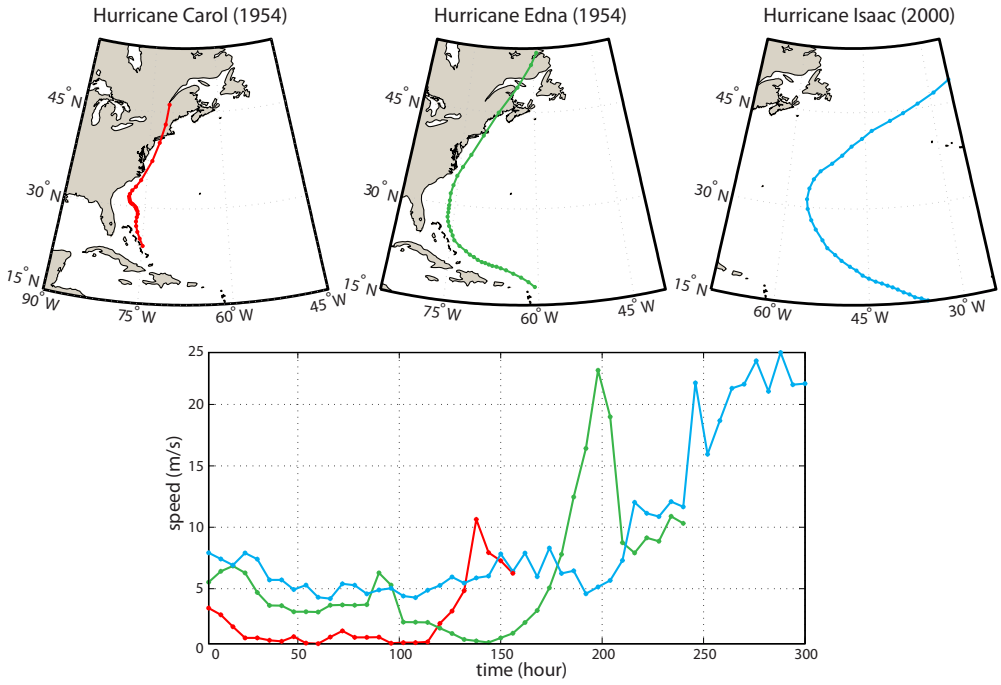


Figure 1. Hurricane Edna (green), exhibited a very similar behavior to her predecessor, Carol (red). Hurricane Isaac (blue) exhibited a different speed pattern and yet similar geometry to hurricane Edna. Data source: http://www.nhc.noaa.gov/

trajectory similarity assessment that relies on trajectory segmentation based on the movement parameters of the objects under study, yielding segments of homogeneous movement characteristics. In this approach, the variability of movement parameters, mirrored in the segmentation of trajectories, is used for expressing trajectory similarity, as opposed to a variety of approaches based on shape-similarities. The segmentation process leads to a simplified, compressed representation of trajectories, called *movement parameter class (MPC) sequence*, which converts movement parameter profiles derived from trajectories into a symbolic representation. In this representation, the important movement characteristics are preserved. We propose a modified version of *edit distance*, termed *Normalized Weighted Edit Distance (NWED)* as the measure of similarity between trajectories. We evaluate NWED in comparison to a relevant similarity measure, called EDM, proposed by Chen *et al.* (2004). Similar to our approach, EDM applies the *edit distance* on *a symbolic representation of trajectories* obtained from the segmentation of two movement parameters, distance and direction between fixes. However, there are distinctive differences in how the movement parameter classes and the edit distance are computed, as will be discussed later.

Additionally, we evaluate our proposed method with two applications in trajectory clusterings: (1) exploiting variability of MP classes, using descriptive statistics of the trajectories' MP classes; and (2) exploiting sequence of MP classes using NWED. These applications are presented as demonstrations to show how our proposed segmentation and similarity analysis techniques can be applied in conjunction with existing standard clustering techniques, such as hierarchical clustering, DBSCAN, or OPTICS (Nanni and Pedreschi 2006), for clustering trajectories according to similarities in the spatio-temporal variation of their movement parameters.

The remainder of this paper is organized as follows: Section 2 gives a structured

overview of previous research on similarity analysis, segmentation of movement data and trajectory clustering. Section 3 introduces the methodology that is proposed in this research. Section 4 describes two application examples of the developed method in trajectory clustering. Section 5 presents three experiments based on two case studies using tracking data of hurricanes and couriers in an urban setting, respectively. Section 6 discusses the key findings of the experiments, the strengths, and limitations of our method in comparison to existing techniques. Finally, Section 7 presents the concluding remarks and outlines directions for future work.

2. State of the Art

2.1. *Movement Similarity Assessment*

Similarity between two objects is quantified as the cost of transforming one entity into another or the distance between the two objects, using a *similarity measure* (Faloutsos *et al.* 1997). So far, a variety of similarity measures has been developed in order to address various aspects of movement similarity assessment problems. The existing movement similarity assessment techniques can be divided into two classes of (1) *spatial similarity*, that is, finding trajectories with similar geometric shape, ignoring the temporal dimension; and (2) *spatio-temporal similarity*, focusing on both spatial and temporal aspects of movement data. Up to now, the proposed movement similarity assessment methods mostly originate from either *time series similarity measures* such as Euclidean distance, edit distance, Longest Common Subsequence (LCSS), and Dynamic Time Warping (DTW) (Ding *et al.* 2008b), or *geometric shape matching* techniques such as Fréchet distance or Hausdorff distance (Alt 2009).

2.1.1. Spatial Measures of Movement Similarity

Most of the recent works on similarity search in trajectory data address the spatial similarity problem (Vlachos *et al.* 2002a,b, Yanagisawa *et al.* 2003, Chen *et al.* 2005, Lin and Su 2005). Euclidean distance based approaches are usually less complex than the other methods and only work on trajectories of the same duration and granularity. The efficiency of such methods decreases when the movement data contain noise, outliers or gaps. The proposed techniques based on LCSS or edit distance are more robust in this respect. The latter approaches can be applied for trajectories of different durations or granularity, albeit at higher computational cost.

The above techniques use the trajectory representation of movement. In contrast, some methods have been proposed that represent trajectories using movement parameters. For example, Little and Gu (2001) apply DTW on separate *path* and *speed* curves of trajectories. In their approach, in order to simplify the process of similarity analysis, a local geometric feature extraction technique is applied using *curvature* information of the path and speed curves. Curvature is invariant to scaling and rigid motion transformations. Another DTW based approach proposed by Vlachos *et al.* (2004) to find similar trajectories of the same granularity under translation, scaling, and rotation transformations. In this approach, a different representation of trajectories based on *turning angle* and *distance* of trajectory fixes over time is applied. DTW based methods allow trajectories to be stretched or compressed and hence do not preserve relative speed of the trajectories. With a different approach, Chen *et al.* (2004) introduced a new representation of trajectories, called *movement pattern strings (MPS)*, in order to optimize similarity computation using an extension of the edit distance, called *Edit Distance on Movement*

Pattern Strings (EDM). MPS is a symbolic representation of a trajectory using *movement direction*, and *distance ratio* information derived from the original trajectory.

The latter methods are relevant to our proposed approach in terms of incorporating *movement parameters* in similarity assessment of moving objects. However, these measures mainly look into the spatial aspect of movement data by taking geometric movement parameters such as distance and direction. Among the aforementioned techniques, the method by Chen *et al.* (2004) is directly comparable to our method since it also applies an *edit distance* on a *symbolic representation of trajectories*. In contrast, Little and Gu (2001) and Vlachos *et al.* (2004) use DTW as similarity measure on metric (non-symbolic) values of movement parameters. Moreover, since Chen *et al.* (2004) uses the relative movement parameters (i.e. distance and direction between two consecutive fixes), the temporal dimension of movement is implicitly involved in the similarity computation when data is sampled at a regular sampling rate. As it is experimentally shown in Chen *et al.* (2004, 2005), and Ding *et al.* (2008b), the edit distance is more accurate than the other commonly used trajectory similarity measures such as Euclidean distance, DTW, LCSS, particularly in the presence of noise in the data. Hence, Experiment #2 in Section 5.1.3 will be devoted to empirically comparing NWED to the method by Chen *et al.* (2004).

2.1.2. Spatio-temporal Measures of Movement Similarity

As pioneering studies in the context of spatio-temporal similarity, Sinha and Mark (2005), Frentzos *et al.* (2007) employed the Euclidean distance for regularly sampled trajectories. Later, van Kreveld and Luo (2007), Buchin *et al.* (2009) improved such techniques to extract the most similar subtrajectories using an approximation from a set of trajectories with different granularity. Pelekis *et al.* (2007) considered a slightly different approach and proposed a family of distance measures, *Locality in In-between Polylines (LIP)*. LIP relies on the *area* of the polygons formed between the intersection points created by the overlay of two trajectories. In order to compute the spatio-temporal similarity between trajectories, different weight factors are applied to support the detection of concurrent movement of objects that move closely at similar speeds, or directions (Pelekis *et al.* 2007). The accuracy of LIP based measures is influenced by penalty factors that need to be specified by user. On the other hand, because the fundamental element of this approach is the area between intersection points, these measures work better for trajectories which follow the same route and are not appropriate for winding trajectories with a lot of turns. Moreover, an additional search process is required to find the intersection points prior to the similarity assessment. Another study by Trajcevski *et al.* (2007) suggested a Rigid Transformation Similarity Distance employing the notion of Fréchet distance to compute the similarity between trajectories under translation and rotation transformations. In a similar attempt, Ding *et al.* (2008a) proposed the *w-constrained Fréchet distance (wDF)*, which constrains the discrete Fréchet distance by a given temporal threshold. The computational cost of these methods is rather high, especially for long trajectories. These measures are not restricted to similar geometric routes, however, they do not consider the speed of the objects either.

2.2. *Trajectory Segmentation*

Segmentation is essential in many applications where the subject of the analysis is a complex and heterogeneous phenomenon (e.g. map generalization, time series analysis etc.). In the study of movement, segmentation facilitates finding patterns and structures

in movement data and hence can help to understand the behavior of objects. Trajectory segmentation refers to decomposing a trajectory into segments of homogenous characteristics. Segmentation has recently been applied in several studies in the domain of moving object data analysis in order to simplify trajectories for several purposes, such as indexing and efficient data handling (Anagnostopoulos *et al.* 2006), event and activity recognition along the geospatial lifelines of objects (Yan *et al.* 2010), and classification of movement data (Dodge *et al.* 2009).

In this study, we use segmentation in order to convert trajectories into revealing structures for extracting similarities in the movement characteristics of objects. Therefore, we extend the segmentation method proposed in an earlier paper by Dodge *et al.* (2009). Their method applies feature extraction techniques from map generalization and time series analysis in order to decompose trajectories into sequences of homogeneous movement characteristics. Buchin *et al.* (2010) recently proposed a similar segmentation approach, however, applying different criteria on movement parameters of objects (e.g. using ranges of speed or turning angle) on a continuous representation of trajectories.

2.3. *Trajectory clustering*

Trajectory clustering is an exploratory data mining technique that, similar to segmentation, facilitates studying movement data and understanding their structure by reducing its complexity. Miller and Han (2009) and Kisilevich *et al.* (2010) provide a survey of the well-known clustering algorithms. Based on these surveys two main classes of clustering techniques used or proposed for movement data include: (a) distance-based clustering techniques, where a distance function (i.e. a similarity measure) is required to determine how the trajectories are grouped together (e.g. k-means, hierarchical clustering, BIRCH); and (b) density-based clustering approaches, where a density threshold is used to identify groups of similar trajectories around an object (e.g. DBSCAN, OPTICS, TOPTICS, TRACLUS). The density-based approaches often treat movement data as point clouds in a space-time cube and cluster points based on their spatial density over time. The similarity measures described in Section 2.1 can be used as a base function in the distance-based clustering approaches to cluster trajectory data in space and time.

A review of the existing approaches suggests that the available trajectory clustering methods are well developed for grouping moving objects according to the closeness of their trajectories in space and time. However, these methods may not be useful to discover clusters of objects with similar variations in their movement parameters, especially when the objects follow different routes. Therefore, in Section 4, we present two examples using our segmentation-based similarity assessment approach for clustering trajectories.

3. Methodology

When an object moves about in space over time (Fig. 2.a), the evolution of its movement parameters can be represented as a function or profile over time (Fig. 2.b). We refer to this function as *movement parameter profile* (or short MP profile). The amplitude and frequency variation of such functions can be quite different for different types of moving objects (Dodge *et al.* 2009). Even more so, the differences may pertain to different episodes of a single object's lifeline, owing to the variability of the underlying physics of movement and the behavior of the object. However, when two or more objects move in a similar way, the corresponding functions will most likely also express similarity.

This has led us to using movement parameter profiles for extracting similarities among trajectories, by decomposing the trajectories into segments (i.e. sections) exposing similar movement characteristics.

An overview of the main methodology used in this work is illustrated in Figure 2 and explained in detail in the subsequent sections. Our approach relies on a discrete trajectory representation (Fig. 2.a), modeling trajectories as a sequence of coordinates over time (Laube *et al.* 2007). The methodology consists of two main processes: (1) trajectory segmentation (Fig. 2.b-c), and (2) similarity computation (Fig. 2.d). The key component of the segmentation algorithm is the movement parameter profile. The segmentation procedure makes use of the amplitude and frequency variations of movement parameters over time and will be explained next. The similarity computation uses a variation of the edit distance as described in Section 3.2.

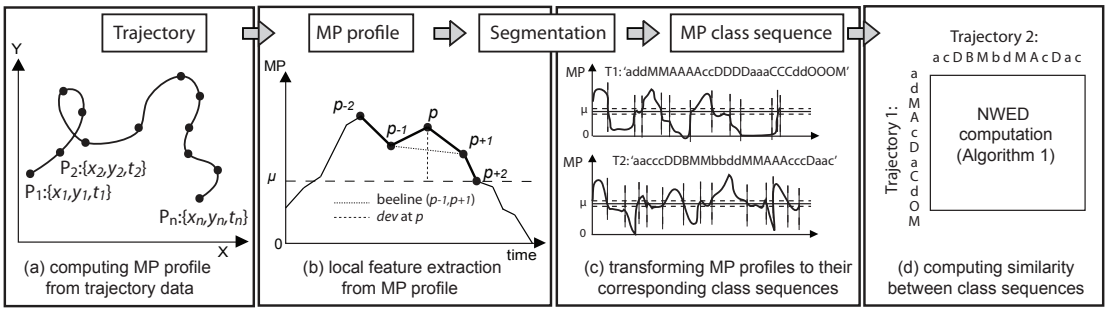


Figure 2. Overview of the trajectory segmentation and similarity computation process.

3.1. *Trajectory Segmentation*

For the purpose of segmentation, an MP profile (e.g. the profile of the movement parameter *speed*) is first generated from trajectory data (Fig. 2.a,b). Then, the MP profile is segmented into sections of homogeneous movement characteristics using an extension of the segmentation approach introduced in Dodge *et al.* (2009). There are several extensions, however, which will be explained in Section 3.1.2. The aim of segmentation is to reduce the complexity of trajectories while conserving their important movement features. That is, instead of using metric values of movement parameters, we characterize movement parameter profiles using a small set of symbols, each being representative of a certain type of variation in movement parameters. This has several advantages: First, segmentation helps to identify salient local features in the movement parameter profiles. Second, it facilitates extracting patterns in the evolution of movement parameters of objects. Third, it enables pattern matching (e.g. similarity analysis) in trajectory data using string matching techniques (Du Mouza *et al.* 2007).

3.1.1. Extracting Local Features from an MP Profile

In order to measure movement characteristics from MP profiles, two measures are used: *Deviation* from the mean value and *sinuosity* of MP profiles, respectively. Deviation gives an impression of the amplitude variation of a movement parameter over time, while sinuosity acts as a proxy of the frequency variation of movement parameters. The amplitude and frequency of variation of an MP profile describe the important features in the evolution function of the corresponding movement parameter over time, and hence identify the main characteristics of movement of an object (Dodge *et al.* 2009).

Figure 2.b provides the supporting illustrations for the computation of the deviation measure and the sinuosity measure on an MP profile. Both measures are defined for each point on the MP profile. The MP profile is first standardized using its standard score ($z = \frac{x-\mu}{\sigma}$) to make it dimensionless and comparable with other profiles. The deviation of a point p on an MP profile equates to its residual value from the mean line (i.e. *dev* in Fig. 2.b). The z-scores indicate the deviation of MP values from the mean line of the original MP profile and therefore have a mean of $\mu = 0$ and standard deviation of $\sigma = 1$. The measure of *sinuosity* for p is computed as a ratio of the distance $\pm k$ points along the profile to the length of the beeline connector centered at the point (i.e. beeline at p for $k = 1$ in Figure 2.b). Where k is the *lag* parameter and is considered as $k = 2$ for GPS observations as discussed. The final sinuosity at p is computed as the average of the computed sinuosity values with different k, as shown in Equation (1). In the end, the sinuosity values are transformed to the interval $[0, 1]$, as proposed in Dodge *et al.* (2009). That is, if profile points are collinear about the given point p the sinuosity measure equals 0 and for a winding profile (i.e. a space-filling curve) it comes to 1.

It should be noted that the lag parameter depends on the temporal granularity, spatial scale, as well as the noise level of the observations. The higher k is set, the bigger the window size gets in the sinuosity computation. And hence, the sinuosity results are smoothed over more points. For instance, for the macro scale hurricane dataset with a temporal granularity of some hours (i.e. 6 hours) and little noise, k can be set to 1. In contrast, for micro scale observations such as eye-tracking data with a temporal granularity of some milliseconds and a high amount of tremors, k can take higher values to reduce the effect of noise.

$$Sinuosity_{p,k} = \frac{\sum_{i=p-k}^{i=p+k-1}(d_{i,i+1})}{d_{p-k,p+k}} \qquad (1)$$

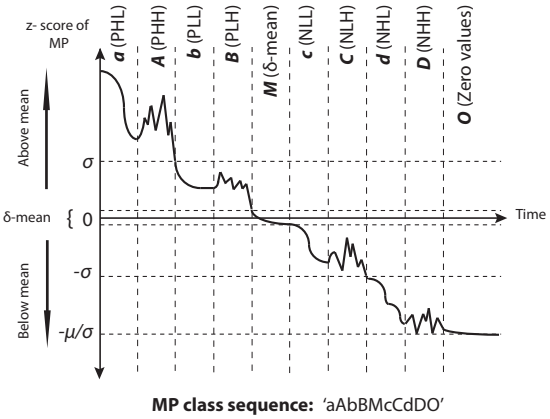
$$Sinuosity_p = \frac{\sum_{j=1}^{j=k} Sinuosity_{p,j}}{|k|}$$

3.1.2. Transforming an MP Profile to an MP Class Sequence

With the movement parameter segmentation procedure, the profile points are classified into two regimes regarding the level of the corresponding sinuosity measure, *low sinuosity* and *high sinuosity*, separated by a user-defined threshold. The same is done with deviation to separate *low deviation* from *high deviation* . In addition to these classes that were introduced in Dodge *et al.* (2009), the position of the points with respect to the mean line is also used to distinguish *positive deviation* (i.e. above the mean) from *negative deviation* (i.e. below the mean). The reason being that since in this study the interest is to detect similarity in the movements of objects over time, it is essential to know whether the amplitude levels of movement parameters are increasing or decreasing. In contrast, Dodge *et al.* (2009) aimed merely at the classification of moving objects according to their intrinsic movement properties. Therefore, the properties of movement parameter profiles were of greater importance than their relative values (e.g. values below or above the mean). Accordingly, the number of *sinuosity* and *deviation* classes is doubled, compared to Dodge *et al.* (2009). Moreover, an additional class of MPs is considered for values within an acceptable error threshold δ from the mean, providing a further facility to deal with noise. Consequently, nine classes are extracted from the segmentation of MP profiles, called *movement parameter class (MPC)* (Fig. 3). *An MPC indicates the type*

of amplitude and frequency variations of a movement parameter (i.e. speed, acceleration etc.).

The number of classes is sought to be small, yet describing the main features of an MP profile. In addition to the aforementioned main movement parameter classes, segments with zero values of MP profiles and with the z-score equal to $\frac{-\mu}{\sigma}$ are tagged as a separate and optional class 'O'. This class obtains importance in application domains like transportation where zero values represent stops (i.e. movement at zero speed) and are treated accordingly.



MPC	Deviation	Sinuosity	Assigned letter
PHH	positive high	high	A
PHL	positive high	low	a
PLH	positive low	high	B
PLL	positive low	low	b
M	δ-mean	—	M
NLH	negative low	high	C
NLL	negative low	low	c
NHH	negative high	high	D
NHL	negative high	low	d

Figure 3. MP classes and MP class sequence.

In this study, the deviation threshold is set to the standard deviation of the z-scores (equal to 1) and the sinuosity threshold is set to 0.80, respectively, following the experience made in Dodge *et al.* (2009). Similarly, the δ threshold, which is the acceptable error threshold, is set to 0.01.

With the segmentation process each trajectory is transformed into a *symbolic representation*, or *string* composed of a sequence of movement parameter classes, called *movement parameter class sequence* or short *class sequence* (Fig. 3). A *movement parameter class sequence* is composed of a string of MP classes representing the transition pattern of movement parameters along a trajectory. The domain of the string is $\Sigma = \{A, a, B, b, C, c, D, d, M\}$. Each character of the string represents a specific class (i.e. A: PHH, a: PHL, B: PLH, b: PLL, C: NLH, c: NLL, D: NHH, d: NHL, and M: δ-mean). This new representation of the trajectories is then employed for the purpose of similarity computation as described in Section 3.2.

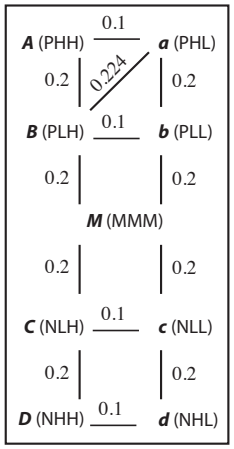
3.2. *Similarity Computation Between MP Class Sequences*

The class sequence representation of MP profiles is now exploited for assessing trajectory similarity. In order to detect similar movement behaviors of objects, our method searches for similar transitions of the MP classes along the trajectories. To do so, the raw trajectories are first transformed to their respective class sequences. Subsequently, in order to calculate the similarity between the sequences we introduce a modified version of the edit distance as a similarity measure, called *Normalized Weighted Edit Distance (NWED)*. The edit distance is used since it is related to the concept of string matching and is a metric to measure the difference between two sequences, in our case MPC sequences. In fact, NWED extends the Levenshtein distance as a well-known form of edit

distance. The Levenshtein distance is defined as the smallest number of insertions, deletions, and substitutions required to convert one string into another (Levenshtein 1966). The edit distance and its variations have been widely used in bioinformatics and speech recognition and recently in similarity analysis of movement data as discussed in Section 2.1.1. Computing the distance between strings with the edit distance has a complexity $O(m \times n)$, where m and n are the lengths of the two strings. Nevertheless, the efficiency of the edit distance can be improved using fast string matching techniques, indexing, or pruning approaches (Du Mouza *et al.* 2007, Ding *et al.* 2008b).

The *Normalized Weighted Edit Distance (NWED)* computes the weighted and normalized cost of converting one MP class sequence into another using *edit operations* (i.e. insertion, deletion, substitution). In comparison to the original edit distance, the modification of NWED concerns the costs of the insertion, deletion, and substitution operators, which all are equal to 1 in the original version of the edit distance. Similar to the original edit distance, the insertion and deletion operations are given the same cost equal to 1. In contrast, we define the cost of substitutions differently for each specific pair of MP classes. That is, the costs are weighted based on the degree of dissimilarity between different classes. The weights vary between 0 (no cost for the same classes) and 1 (maximum cost). Since inserting or deleting points changes the length of MP profiles (i.e. duration of movement), we consider such operations to be more severe than the substitution of different MP classes, hence, the other substitution costs shall be less then these costs. Moreover, as it is illustrated in figures 3 and 4, we assign *four* degrees of differences between the considered *amplitude levels*. The substitution cost of two consecutive amplitude levels is considered equal to 1 (i.e. σ of the z-score values as described in Section 3.1.2). In order to normalize the costs to the domain $[0, 1]$, all costs are then divided by 5 (i.e. 4 amplitude degrees + 1 insertion/deletion cost), leading to 0.2 for each substitution cost of two consecutive amplitude levels. The remaining substitution costs are then computed accordingly as described below.

Traditionally, in time series analysis, long term and short term variations are treated separately (Foster 1996). For that reason, we assign different substitution costs for amplitudes (i.e. deviations) and frequencies (i.e. sinuosities). That is, the substitution cost between *low deviation* and *high deviation* classes is assigned twice the substitution cost than between *low sinuosity* and *high sinuosity* classes. The reason being that the amplitudes give an indication of the long term variation of movement parameters and hence a considerable cost (i.e. time, energy) is required to transit from one level of amplitude to another. In contrast, the frequencies capture the short term variation and local features of the MP profiles, hence, the transition between the two classes of sinuosities at the same level of amplitude involve less cost. Accordingly, MP values of the same level of deviation but with a different sinuosity regime are considered more similar than the ones that deviate more yet have the same sinuosity regime. For instance, MP values of class $'a'$ are considered more similar to an $'A'$ segment than to a $'b'$ segment (see Figure 3 and Figure 4). By the same token, $'C'$ is more similar to $'c'$ than to $'B'$. Accordingly, the other costs are determined using the Pythagorean theorem as shown in Figure 4. These costs, summarized in $COST\_Matrix$ in Equation (2), indicate the degree of dissimilarity between the different trajectory segments that belong to different MP classes.



$d(x_i, \phi) = 1$ insertion/deletion cost

$d(B, C) = d(b, c) = 2/5 = 0.4$

$d(B, c) = d(b, C) = 2.06/5 = 0.412$

$d(A, D) = d(a, d) = 4/5 = 0.8$

$d(A, d) = d(a, D) = 4.03/5 = 0.806$

$d(A, B) = d(B, M) = d(M, C) = d(C, D) = 1/5 = 0.2$

$d(a, b) = d(b, M) = d(M, c) = d(c, d) = 1/5 = 0.2$

$d(A, a) = d(B, b) = d(C, c) = d(D, d) = 0.5/5 = 0.1$

$d(A, b) = d(B, a) = d(C, d) = d(D, c) = 1.12/5 = 0.224$

$d(A, c) = d(a, C) = d(b, D) = d(B, d) = 3.04/5 = 0.608$

$d(A, C) = d(a, c) = d(b, d) = d(B, D) = 3/5 = 0.6$

$d(A, M) = d(a, M) = d(d, M) = d(D, M) = 2/5 = 0.4$

Figure 4. Costs (i.e. dissimilarities) of the MP classes employed in NWED computed using the Pythagorean theorem.

$$\text{COST\_Matrix} = \begin{array}{c} \\ PHH \\ PHL \\ PLH \\ PLL \\ MMM \\ NLH \\ NLL \\ NHH \\ NHL \end{array} \begin{array}{c} \begin{array}{ccccccccc} PHH & PHL & PLH & PLL & MMM & NLH & NLL & NHH & NHL \end{array} \\ \begin{pmatrix} 0 & 0.1 & 0.2 & 0.224 & 0.4 & 0.6 & 0.608 & 0.8 & 0.806 \\ 0.1 & 0 & 0.224 & 0.2 & 0.4 & 0.608 & 0.6 & 0.806 & 0.8 \\ 0.2 & 0.224 & 0 & 0.1 & 0.2 & 0.4 & 0.412 & 0.6 & 0.608 \\ 0.224 & 0.2 & 0.1 & 0 & 0.2 & 0.412 & 0.4 & 0.608 & 0.6 \\ 0.4 & 0.4 & 0.2 & 0.2 & 0 & 0.2 & 0.2 & 0.4 & 0.4 \\ 0.6 & 0.608 & 0.4 & 0.412 & 0.2 & 0 & 0.1 & 0.2 & 0.224 \\ 0.608 & 0.6 & 0.412 & 0.4 & 0.2 & 0.1 & 0 & 0.224 & 0.2 \\ 0.8 & 0.806 & 0.6 & 0.608 & 0.4 & 0.2 & 0.224 & 0 & 0.1 \\ 0.806 & 0.8 & 0.608 & 0.6 & 0.4 & 0.224 & 0.2 & 0.1 & 0 \end{pmatrix} \end{array} \quad (2)$$

Algorithm **1** presents the computation process of the NWED between two class sequences T and P. In order to compute the dissimilarity between two segmented trajectories (i.e. two MP class sequences), one is considered as *subject trajectory* (i.e. $T[1 \dots n], \quad |T| = n$) and the second one is considered as a *pattern* or *template* (i.e. $P[1 \dots m], \quad |P| = m$). A $n \times m$ dissimilarity matrix (i.e. WED_Matrix) is then formed based on the costs between segments of the two trajectories, obtained from $COST\_Matrix$ (Equation (2)), applying Equation (3a) (Bozkaya *et al.* 1997). The element $WED\_Matrix(n, m)$ indicates the cost of conversion between the two sequences or the dissimilarity between the two trajectories. Finally, in order to remove the effect of varying length of trajectories on the similarity results, we normalize the total dissimilarity between T and P, as proposed by Yujian and Bo (2007) using Equation (3b). The NWED obtained from Equation (3b) is metric, as proven by Yujian and Bo (2007).

$$WED\_Matrix_{T,P} = C_{0 \dots n, 0 \dots m} \quad (3a)$$

$$C_{i,j} = \begin{cases} j & \text{if } i = 0 \\ i & \text{if } j = 0 \\ C_{i-1,j-1} & \text{if } i, j > 0 \text{ and } T_i = P_j \\ COST\_Matrix\,(T_i, P_j) + \min\,(C_{i-1,j-1}, C_{i-1,j}, C_{i,j-1}) & \text{otherwise} \end{cases}$$

$$NWED_{T,P} = \frac{2 \times C_{n,m}}{n + m + C_{n,m}} \quad (3b)$$

Figure 5 provides an example of running the segmentation and NWED algorithms on four hurricane trajectories of different durations (i.e. of unequal length), including hurricanes Carol (H1), Edna (H2), and Isaac (H3) mentioned in the Introduction (Fig. 5.a). Hurricanes Carol and Edna exhibit a rather similar increase-decrease speed pattern in their evolution from formation to decaying after landfall, although their movement paths differ. Hurricane Isaac is the longest and its movement path exhibits a relatively similar curve to hurricane Edna. Contrary to the other selected hurricanes, the speed profile of

Algorithm 1 NWED

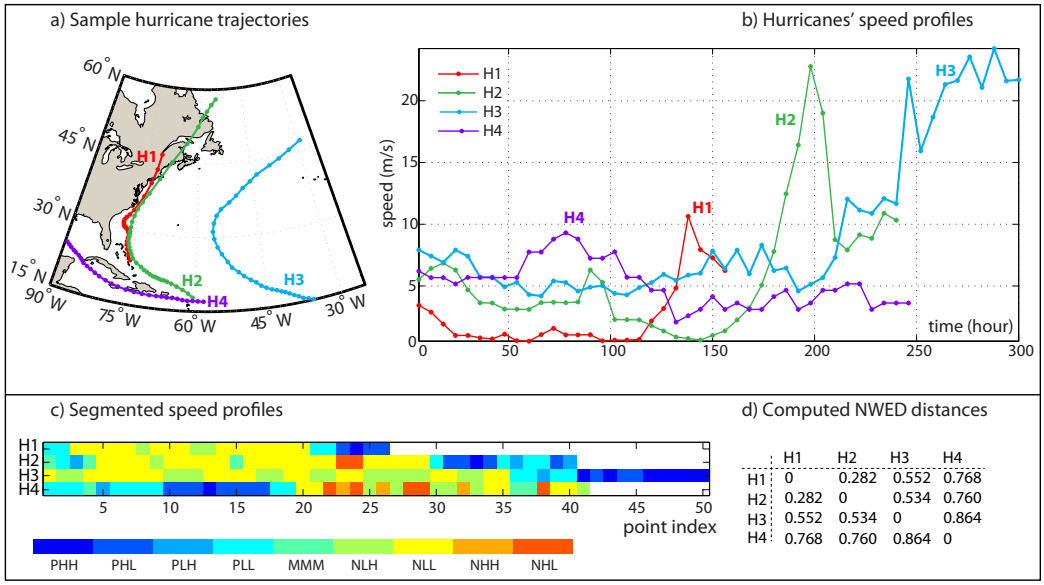
Require: input: two sequences T and P, lengths of T and P
Require: input: $COST\_Matrix$ (weighted costs between MP classes)
Ensure: output: NWED between two sequences T and P

$n \leftarrow |T|$
$m \leftarrow |P|$
$sumLength \leftarrow (n + m)$
for $i = 0$ to n **do**
  for $j = 0$ to m **do**
    if $i = 0$ **then**
      $WED\_Matrix(i, j) \leftarrow j$
    else if $j = 0$ **then**
      $WED\_Matrix(i, j) \leftarrow i$
    else if $T_i = P_j$ **then**
      $WED\_Matrix(i, j) \leftarrow WED\_Matrix(i - 1, j - 1)$
    else
      $mpcCOST \leftarrow COST\_Matrix(T - i, P - j)$
      $minCOST \leftarrow \min \begin{cases} WED\_Matrix(i - 1, j - 1) \\ WED\_Matrix(i - 1, j) \\ WED\_Matrix(i, j - 1) \end{cases}$
      $NWED\_Matrix(i, j) \leftarrow mpcCOST + minCOST$
    end if
  end for
end for
$WED_{T,P} \leftarrow WED\_Matrix(n, m)$
$NWED_{T,P} \leftarrow \frac{2 \times WED_{T,P}}{sumLength + WED_{T,P}}$
return $NWED_{T,P}$

Isaac (H3) shows a variable, yet overall an increasing pattern, since the hurricane does not make landfall. On the other hand, hurricane H4 (San Zacarias, 1910) moves along a relatively straight path but its speed varies more frequently. As the figure illustrates, first of all the speed profiles are computed from trajectories (Fig. 5.b). Then, the speed profiles are converted into their corresponding class sequences using the segmentation method (Fig. 5.c). Figure 5.d presents the computed pairwise NWED between the four hurricanes. As it can be seen from the figure, the NWED between hurricanes Edna (H1) and Carol (H2) is computed as 0.282, which is smaller with respect to the other pairwise NWED distances, since both hurricanes show a relatively similar sequence in their speed variations. The computed NWED between Isaac (H3) and Edna (H2), 0.534, indicates that these two hurricanes are relatively dissimilar in terms of their speed patterns. Hurricane H4 is obtained as the most dissimilar hurricane ($NWED > 0.7$), since its speed exhibits a distinctively different pattern with respect to the other hurricanes.

4. Application: Trajectory Clustering

The proposed methodology can be applied in different application domains, whenever the aim is to discover common behaviors of dynamic entities in space and time. These include applications such as trajectory clustering (e.g. grouping trajectories with similar speed patterns) and movement pattern detection (e.g. discovering concurrence patterns). Below, we present two examples where we apply our segmentation-based similarity approach for clustering trajectories. Both examples use standard distance-based clustering approaches. The first example is intended to show how the proposed segmentation tech-



	H1	H2	H3	H4
H1	0	0.282	0.552	0.768
H2	0.282	0	0.534	0.760
H3	0.552	0.534	0	0.864
H4	0.768	0.760	0.864	0

Figure 5. Computation of NWED on four hurricane trajectories including hurricanes Carol (H1), Edna (H2), and Isaac (H3).

nique helps grouping trajectories with similar variability in the characteristics of their movement parameters. In contrast, the second example demonstrates how the proposed similarity measure can be exploited in grouping trajectories with similar pattern sequences in the evolution of their movement parameters. In both examples, segmentation enables capturing important features in the variations of movement parameters over time, rather than simply clustering the trajectories based on geometric properties. However, a filtering process such as the one presented in Etienne *et al.* (2010) can be applied as a preprocessing step in order to ensure route similarity. Both examples can in principle be applied in conjunction with any standard clustering techniques such as hierarchical clustering, K-means, or DBSCAN. However, clustering results may vary according to the clustering techniques. In this study, we applied the complete-linkage agglomerative hierarchical clustering technique (Miller and Han 2009).

4.1. *Exploiting Variability of MP Classes in Trajectory Clustering*

The first clustering example is based on descriptive statistics computed on the MP classes resulting from the segmentation process. For each MP, the number of transitions of different classes and the percentage, standard deviation and average length of each MPC of the domain (i.e. $\Sigma = \{A, a, B, b, C, c, D, d, M\}$) in the MPC sequence is computed. The number of transitions of each class gives an indication of the variability in the dynamic trend of the MP and the percentage of each class in the MP profile indicates the frequency of the contribution of each class of Σ to the total trend of the MPC sequence. Therefore, for each segmented trajectory the following features are computed: percentage of the number of class alterations (i.e. 1 feature), percentage of contribution of each class (i.e. 9 features), the mean and standard deviation of each class length (i.e. 18 features). Thus, the total number of features that are fed to the clustering process is 28. Following that, a standard hierarchical clustering approach can be used to cluster the segmented trajectories (i.e. sequences) from the computed features. This application of clustering focuses more on the *variability* of the movement parameters rather than the sequence of

the transitions. Therefore, we refer to it as *MPC variability-based similarity.*

4.2. *Exploiting Sequence of MP Classes in Trajectory Clustering*

The second example uses the NWED distance function introduced in section 3.2 for similarity-based clustering of trajectory data. Hence, a distance matrix is computed here from the pairwise NWED between segmented trajectories (i.e. sequences). This distance matrix is then used as an input for the clustering process, along with the complete-linkage agglomerative hierarchical clustering to group trajectories with similar trend in the transition of MP variations. In comparison to the previous application (Section 4.1), which considers more the variability of the segments, this application of clustering focuses on the transition *sequence* of the classes of the MP profiles. Therefore, we refer to it as *MPC sequence-based similarity.* This clustering example is recommended for studies where the evolution of the movement parameters is important, such as movement behavior study of hurricanes, and homing pigeons (Laube *et al.* 2007).

5. Experiments

We conducted two case studies in order to assess the applicability of the proposed methods. Thereby, we evaluate our segmentation-based similarity measure with two examples where trajectories are clustered based on our similarity. In these two case studies, the proposed methods were applied on two distinct types of movement data, from different application domains and with rather different dynamic behaviors. From the domain of meteorology, we considered tracks of North Atlantic hurricanes, which express rather smooth and predictable trajectories. In contrast, from the transportation domain, GPS trajectories of couriers were analyzed. The latter dataset involves very diverse dynamic behaviors. Additionally, we compared the outcomes of our similarity measure with the one introduced by Chen *et al.* (2004). To do so, we conducted a comparative experiment employing hurricane tracking data since we have the background information from meteorological literature to validate the computed similarities. Table 1 summarizes the objectives of the conducted experiments.

Table 1. Overview of the Experiments

Case Study	**Data**	**Exp. No**	**Objective**
I	Hurricanes	Exp. #1	Assessment of the MPC variability-based versus sequence-based similarity in seeking structures in hurricanes.
		Exp. #2	Evaluation of NWED in comparison to the method by Chen *et al.* (2004) (i.e. EDM) on hurricanes.
II	Couriers	Exp. #3	Assessment of the MPC variability-based versus sequence-based similarity in exploring traffic patterns.

5.1. *Case Study I - Clustering North Atlantic Hurricanes*

According to meteorological studies, a hurricane develops gradually in different phases, from formation to decay after landfall (Elsner and Kara 1999). During each phase of the

lifecycle, hurricanes to a great extent have similar characteristics. Apart from meteorological prerequisites, the *season* and the *geographic latitude* of the hurricanes' origin are two important factors influencing the dynamic behavior of hurricanes. Elsner and Kara (1999) distinguish between 'classic' *low-latitude* hurricanes originating south of about 20° N and *high-latitude* hurricanes originating north of about 20° N. Furthermore, the authors state that North Atlantic Hurricanes have similar dynamic characteristics based on the time of formation and their source locations. That is, it is observed that hurricanes of similar season (i.e. fall or summer) with similar characteristics usually originate in a spatial proximity (i.e. in the same quadrants w.r.t. (19° N, 80° W)). This case study aims at evaluating our similarity assessment and clustering approaches by confirming these findings. Therefore, we primarily sought for extracting two clusters in the historical tracks of hurricanes to see whether the results correspond with the categorization of the *low-latitude* and *high-latitude* hurricanes. Moreover, we were interested to assess whether the clusters relate to the season in which hurricanes occurred or to the distance of their origin to the US coastline.

5.1.1. Data

Since destructions caused by hurricanes mainly happen after landfall, it is most important to investigate the dynamic behavior of hurricanes that reach the coastline. Hence, this case study uses trajectories of 397 hurricanes that made landfall at the East or South East coastline of the United States between 1907 and 2007[1]. The temporal sampling rate of the observations is 6 hours. Since the raw hurricane movement dataset obtained from NOAA contains little noise and is regularly sampled, no preprocessing was required prior to the segmentation and clustering procedures.

5.1.2. Experiment #1: Assessment of MPC Variability-Based vs. Sequence-Based Similarity in Clustering

The aim of this experiment is to evaluate the usefulness of the proposed segmentation-based similarity assessment approach for seeking structure in a movement dataset with sequential behavior. The similarity of hurricane trajectories was assessed based on their *speed* and *turning angle* (change of movement direction) behavior, since such MPs are key parameters that affect the time and location of a hurricane' landfall.

For clustering exploiting MPC variability-based similarity (cf. section 4.1), 28 descriptive statistics features were derived for 397 trajectories, for both segmented MP profiles, and used for hierarchical clustering. Two clusters for speed profiles and two clusters for turning angle profiles were generated. The clustered MP profiles are presented in Figure 6. As Figures 6.a illustrates, both speed clusters exhibit a relatively smooth transition from NLL (yellow) to PHH (dark blue). Similarly, a gradual transition from MMM (aquamarine) to PHL (light blue) can be observed in the turning angle profiles (Figure 6.b). Compared to the speed profiles, more variability can be observed in the turning angle behavior.

For clustering exploiting MPC sequence-based similarity (cf. section 4.2), NWED is used to compute distance matrices between segmented MP profiles. Two distance matrices of extent 397×397 were computed, one for speed and one for turning angle, which were then used separately for clustering. Two clusters were generated for both speed and turning angle profiles. Figure 7 illustrates the map view and the class sequence representations of the two clusters for speed only. The obtained clusters also reveal the gradual

[1] from NOAA's Coastal Services Center (http://csc-s-maps-q.csc.noaa.gov/hurricanes/)

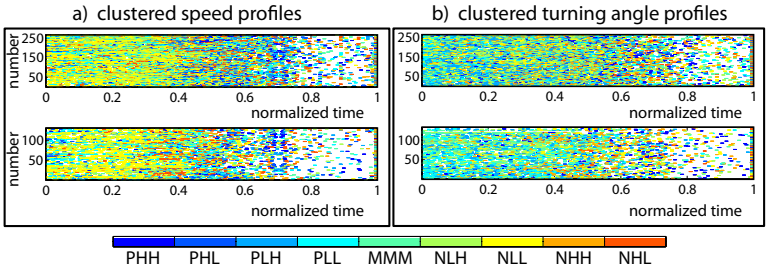


Figure 6. Experiment #1 (exploiting variability-based similarity): Class sequences of the two clusters of a) speed, and b) turning angle.

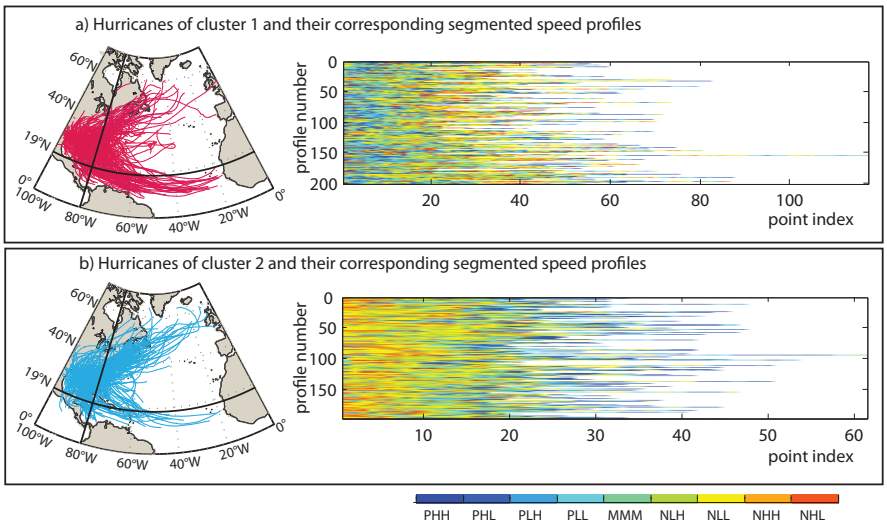


Figure 7. Experiment #1 (exploiting sequence-based similarity of speed profiles): The map view and segmented speed profiles of the two clusters of North Atlantic Hurricane trajectories.

increase-decrease pattern in the speeds of hurricanes. However, the trend of this pattern varies in the two clusters: From NLH over NLL to PHH in cluster 1 (Figure 7.a), and from NLL to PHL in cluster 2 (Figure 7.b).

In order to test whether clusters reveal significant difference between hurricanes of different months or hurricanes originating from different locations, we applied Mann-Whitney U tests. The resulting p-values (see Table 2, Exp. #1) show which attributes do or do not explain the two generated clusters: latitude of origin, longitude of origin, low-latitude vs. high-latitude of origin, origin west or east of 80° W, season (fall or summer), and month of the year. The results indicate no significant difference in the source location of hurricanes in the obtained clusters from variability-based similarity (p-values above 0.05). Since the hurricanes generally show a gradual increasing-decreasing speed trend (speeding up after formation, slowing down after landfall) (Elsner and Kara 1999), their speed profiles exhibit little variability and rather a sequential behavior. Therefore, the effectiveness of the variability-based similarity decreases for clustering such sequential behavior. In contrast, the clusters obtained applying sequence-based similarity reveal a significant difference between trajectories originating from the latitudes north and south of 19° N ($p < .001$, Table 2, Exp. #1; Figure 7). Hence, exploiting sequence-based similarity (using NWED) for clustering was capable of distinguishing between *low-latitude* and *high-latitude* hurricanes. Moreover, with respect to longitude, clusters obtained using sequence-based similarity differentiate hurricanes originating east and

Table 2. Experiment #1: Computed p-values of the Mann-Whitney U test on the clustering results for hurricanes (* $p < 0.05$)

Exp.	**Method**		**No. of clusters**	**Lat. of origin**	**Long. of origin**	**Low vs. high Lat. origin**	**Origin west or east of 80°W**	**Season**
#1	variability-based	speed	2	0.107	0.356	0.249	0.068	0.019*
		turning angle	2	0.354	0.323	0.088	0.187	0.497
	NWED	speed	2	.000*	.000*	.000*	.000*	0.001*
		turning angle	2	.000*	.000*	.000*	.000*	0.196
#2	EDM		2	.000*	.000*	.000*	0.025*	0.281

west of 80° W ($p < .001$, Table 2, Exp. #1). This distinction reflects the distance of the hurricanes' origin to the US coastline (see map view in Figure 7).

On the other hand, the clusters obtained from both variability-based and sequence-based similarity on speed profiles suggest a significant difference between hurricanes in different seasons (i.e. summer and fall) ($p < 0.05$ in Table 2, Exp.#1). In cluster 1 obtained from sequence-based similarity, we find a tendency for hurricanes in summer (i.e. May to August) and originating from the southeastern quadrant of (19° N, 80° W) (Figure 8.a). In contrast, cluster 2 predominantly shows hurricanes in fall (i.e. September to December), originating from the northwestern quadrant of (19° N, 80° W) (Figure 8.b). This observation complies with results from the meteorology literature (Elsner and Kara 1999).

The clusters obtained from turning angle profiles suggest similar outcomes (Table 2, Exp. #1). The result of exploiting sequence-based similarity for clustering suggests that hurricanes that originate in the same region (quadrant) to a certain extent tend to follow a similar change in their direction or geometric shape (i.e. spatial similarity). However, the trend of this variations does not significantly differ in time (i.e $p = 0.196$ for different seasons). In contrast, we could not find such structure from the results of variability-based similarity of the turning angles profiles ($p > 0.05$).

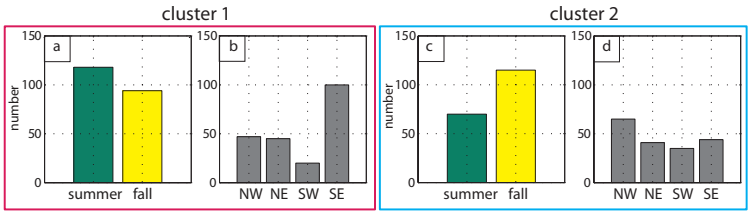


Figure 8. Experiment #1 (exploiting sequence-based similarity of speed profiles): Properties of resulting two clusters. Histograms of clusters formed, by season and source location of hurricanes. Source locations are given as quadrants NW, NE, SW, and SE of 19° N latitude and 80° W longitude.

5.1.3. Experiment #2: Comparing NWED and EDM

This experiment compares NWED with EDM introduced by Chen *et al.* (2004). For this comparative study we implemented the EDM similarity measure (described in Section 2.1.1), which computes the spatial similarity between *movement pattern string (MPS)*

representations of trajectories (Chen *et al.* 2004). The main motivation for this comparative study is that EDM is one of the few available techniques similar to our approach. In detail, EDM and NWED are comparable as they share the following specifications:

- Both consider movement parameter informations in trajectory similarity assessment (e.g. movement direction);
- both measures are an extension of edit distance, and compute distance between symbolic representations of trajectories; and
- both have a similar computational complexity (i.e. $O(n^2)$).

Chen *et al.* (2005) have already investigated the performance of edit distance in comparison with the other similarity measures such as LCSS, DTW, and Euclidean distance. Their study suggests that edit distance is more accurate and robust, specially in the presence of noise and time lags between similar trajectories (local time shifts). Therefore, here we specifically only compare the similarity results of NWED (i.e. sequence-based similarity from Experiment #1) to EDM on clustering the hurricane trajectories, but not the computational complexity. However, in order to make both measures comparable, we normalized EDM to the scale $[0, 1]$ as it is described for NWED in Section 3.2. As recommended, we used a distance threshold of $\epsilon - dis = 0.125$ and a direction threshold $\epsilon - dir = \pi/4$ to generate the 8×8 (movement direction and movement distance) quantization map (Chen *et al.* 2004). Next, movement pattern string (MPS) sequences had to be derived for all hurricane trajectories using the quantization map (Chen *et al.* 2004). Finally, a 397×397 distance matrix for EDM was computed. Just as in Experiment #1, the EDM distance matrix was used for trajectory clustering (again complete-linkage). It is necessary to remark that since hurricane data are sampled at a regular interval, *movement distance* in EDM implicitly represents the speed information of hurricanes, and hence, EDM is indeed comparable to our NWED measure for this case study.

As in Experiment #1 two distinct clusters were generated based on EDM. We applied a Mann-Whitney U test on the clusters, as is described in Experiment #1. The resulting p-values suggest that (see Table 2): The two clusters show significant differences on all attributes related to the location of hurricanes' origin (Lat. and Long. of origin; low vs. high Lat. origin; origin west or east of 80°W). By contrast, the clusters do not significantly differ regarding time-related attributes (i.e $p = 0.281$, and $p = 0.345$ for different seasons and months, respectively). These findings are similar to the results from the clusters obtained from turning angle based on NWED (Experiment #1). The reason being that both MPS and turning angle sequences capture the geometric shape of the hurricanes.

Furthermore, we aimed at examining the difference between distances obtained from EDM in comparison to NWED (based on speed). To do so, we first grouped all hurricanes based on the time of formation (i.e. fall or summer), and then based on the location of their origin (i.e. northwestern and southeastern quadrants of (19° N, 80° W)). We then computed descriptive statistics for the NWED and EDM distance distributions for the two groupings (Figure 9 and Table 3). We applied the Mann-Whitney U test on the histograms of each group separately, in order to see if the histograms were the same or different. The resulting p-values (i.e. $p < .001$) indicate for all four cases that the distribution of the NWED and EDM significantly differs (Figure 9 and Table 3). Moreover, having a closer look at the histograms of NWED, we could infer that the hurricanes originating in a spatial and temporal proximity of each other (hurricanes of each group) exhibit a similar speed behavior (mean NWED distances ≤ 0.3). This observation confirms the hypothesis of this case study. By contrast, the results obtained from EDM distances do not reveal such similarity (mean EDM distances > 0.5). Hence,

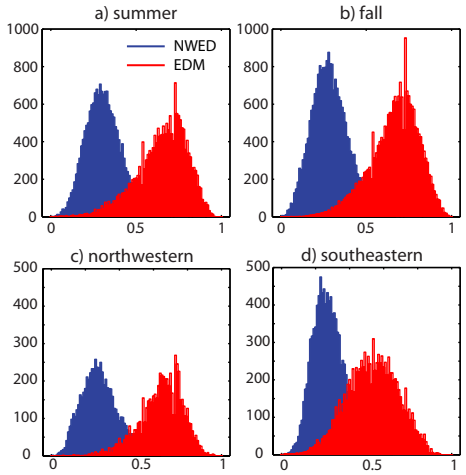


Figure 9. Experiment #2: Histograms of the NWED and EDM distance matrices of hurricanes in a) summer, b) fall, c) northwestern of (19°N, 80°W), and d) southeastern of (19°N, 80°W)

Table 3. Experiment #2: Descriptive statistics and Mann-Whitney U test on NWED and EDM distances

Groups	**No. of distances in groups**	**NWED**				**EDM**				*p*-**values**
		Mean	**Std.**	**Median**	**Skewness**	**Mean**	**Std.**	**Median**	**Skewness**	
summer	17578	0.30	0.10	0.3	0.29	0.65	0.14	0.67	-0.71	.000
fall	21736	0.28	0.10	0.27	0.37	0.66	0.14	0.68	-0.63	.000
northwest	6216	0.28	0.11	0.27	0.36	0.64	0.13	0.66	-0.68	.000
southeast	10298	0.27	0.09	0.27	0.47	0.53	0.15	0.53	-0.15	.000

EDM seems to be insufficient in studying the similarity of the speed patterns of hurricanes since it can only capture the spatial similarity.

5.2. *Case Study II - Clustering Courier Trajectories*

In traffic management, it is important to understand the traffic patterns on a given street network over space and time. Quantifying the similarity of vehicles moving on a specific section of a street network can help to distinguish normal and abnormal traffic patterns. This second case study is based on trajectories of couriers captured in Central London by the eCourier company[1] during the month of November 2009. The proposed similarity measures shall be used to discover traffic patterns of vehicles moving on a particular section of the street network, based on their speed behaviors.

5.2.1. Data

The raw GPS data have a temporal sampling rate of approximately one fix per 10 seconds. Two subsets were extracted from all courier trajectories. The first subset, named *ZoneData*, covers the Congestion Zone of London[2]. The aim was to evaluate the performance of our approach on a large transportation dataset with diverse behaviors. The second subset, named *RouteData*, contains sets of trajectories that follow a given route.

[1]http://www.ecourier.co.uk
[2]http://www.tfl.gov.uk/tfl/roadusers/congestioncharge/whereandwhen/

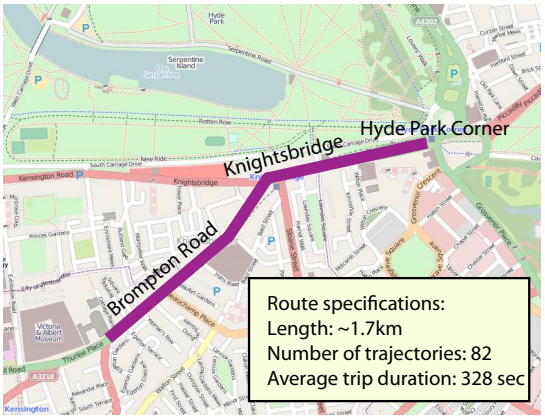


Figure 10. The selected route of courier trajectories (basemap: OpenStreetMap.org)

The selected route leads from Hyde Park Corner to the end of Brompton Road, via Knightsbridge (Figure 10). The reason to restrict the study area to a specific route is first to remove the effect of the geometric shape of the road network on the similarity computation from trajectories; and second to render the courier trajectories comparable to the hurricane trajectories (which have a relatively similar geometric shape).

The courier data required an elaborate pre-processing procedure including various filtering and resampling techniques. Here, however, we only list the most important pre-processing steps. First, outliers (speed over 20 ms^{-1}) and stops were removed from the raw GPS tracking data. Next, stops (speed below 1 ms^{-1}) representing deliveries or stops at traffic lights were filtered from the remaining data as suggested in Doherty *et al.* (2001), since stops can contain errors due to loss of signal. Finally, the cleaned trajectories were resampled using linear interpolation to achieve trajectories with a temporal granularity of exactly 10 seconds. Since the raw trajectories contain information about temporal properties and movement parameters, in order to maintain that information no additional smoothing or map matching that could change the geometry of the trajectories was applied.

The average movement phase between two deliveries is approximately 15 minutes. For that reason, for the *ZoneData* we partitioned the pre-processed trajectories into subtrajectories of 15 minutes duration. Eventually, 100 random samples of *small van* trajectories per hour between 8 AM to 8 PM during weekdays were selected (i.e. a total of 1200 subtrajectories of 15 min duration). For *RouteData*, subtrajectories that followed the aforementioned selected route were then extracted from the entire pre-processed dataset applying geometric curve matching within a threshold distance of 30 m. Overall, a total of 71 trajectories (i.e. 35 motorbike and 36 vans) with an average duration of 323 seconds were obtained on the selected route between 8 AM to 20 PM during weekdays.

5.2.2. Experiment #3: Exploring Traffic Patterns over Time

Following the pre-processing procedure, the speed profiles speed were derived for both *ZoneData* and *RouteData*. The segmentation and clustering process was applied on the generated profiles, just as in Experiment #1.

First, we aimed at testing the applicability of our approach for finding trajectories with similar speed patterns along a specific route, where the effect of geometry is limited. We clustered the trajectories applying sequence-based similarity using NWED on the speed profiles of courier trajectories from *RouteData*. Two clusters were generated. The clustered, segmented speed profiles of courier trajectories as well as the space-time cube

representation of the corresponding trajectories are illustrated in Figure 11. As it can be seen from the segmented profiles in Figure 11.b, the two obtained clusters show different speed behaviors of the courier vehicles. Specifically, the first cluster represents fast movements with a smooth transition from NLL (yellow) to PHH classes (dark blue). In contrast, the second cluster features diverse movement behaviors, including stop-and-go movements (i.e. repetitive sequences of PLL-NLL-NHH-NHL classes) at the beginning of the trajectory (Figure 11.c). The movement behavior represented by the obtained clusters to a great extent reveals the major traffic light about midway on the route (see Figure 11.a,c). The results suggest that exploiting MPC sequence-based similarity in clustering helps to detect the traffic behavior along particular segment of a street network.

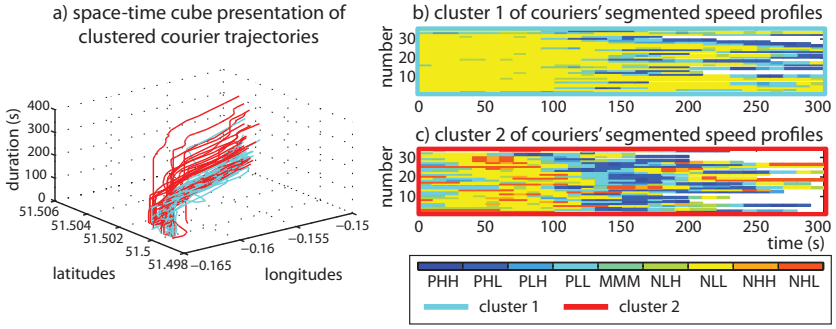


Figure 11. Experiment #3 (exploiting sequence-based similarity, RouteData, speed): a) space-time plot, z-axis represents duration, all trajectories are synchronized to start at time 0 s. b) speed class sequences of the resulting clusters

Next, we applied both variability-based and sequence-based similarities on *ZoneData*, which contain a larger extents of the street network in comparison to the *RouteData*. The aim was to test whether our methods can detect clusters corresponding to the traffic peak and off-peak hours using speed profiles of the couriers. We were interested to find diurnal time windows corresponding to three categories of *low speed* (i.e. slow movements), *medium speed* (i.e. moderate movements), and *high speed* (smooth movements). Therefore, three clusters were generated from the segmented speed profiles of *ZoneData* using the variability-based similarity (Figure 12). We used the Kruskal-Wallis test in order to first test whether clusters are significantly different in terms of the mean speed. And second, to test whether clusters represent specific diurnal periods. Both tests were resulted in $p < .001$. Hence, the results confirmed that the mean speed of 3 clusters are significantly different ($p < .001$, see box-and-whisker plots in Figure 12.a), although we did not use mean speed as a feature in the clustering. Furthermore, the diurnal composition significantly differs between the three clusters ($p < .001$, see Figure 12.b-d). As it can be seen in Figure 12, cluster 1, which represents very slow movement (mean = 3.79 ms^{-1}), has higher frequency during the morning peak (8 - 10 AM) and in the afternoon (3 - 5 PM). In contrast, cluster 2 with medium speed (mean = 4.72 ms^{-1}) represents trajectories at noon (11 AM to 2 PM) and in the evenings (after 6 PM). Cluster 3 with mean speed of 5.7 ms^{-1} did not represent any specific time period. However, it shows a slightly higher percentage in the evenings. These results to a some extent are comparable to the available traffic information about the London Congestion Zone (i.e. AM peak (7 - 10 AM), inter-peak (10 AM - 4 PM), and PM peak (4 - 7 PM))[1].

[1]Transport for London (http://www.tfl.gov.uk/)

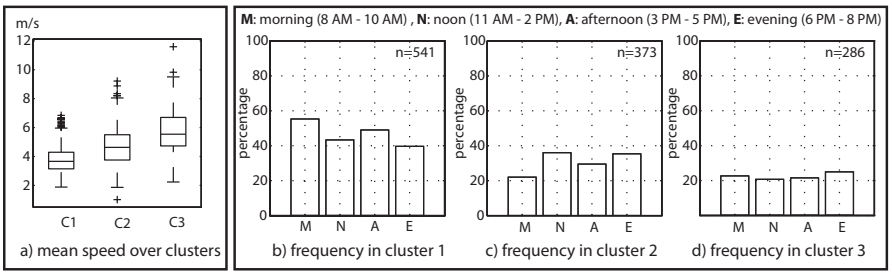


Figure 12. Experiment #3 (exploiting variability-based similarity of speed profiles, ZoneData): Mean speed and frequencies per hour of van trajectories in the three clusters

Just as in Experiment #1, we also applied clustering on the pair-wise NWED distances between segmented speed profiles of van trajectories for *ZoneData*. However, we could not relate the obtained clusters to any specific time periods. This can be explained with the fact that the *ZoneData* trajectories were obtained on a street network heterogeneous with respect to geometry and traffic. Therefore, the interpretation of a sequential movement behavior would not be valid for such networks. In addition, in the contrary to the *RouteData*, we did not have any information about the geographic context of the trajectories (i.e. traffic lights, delivery locations, type of street etc.) in order to make plausible assumptions to validate the results.

6. Discussion

Based on the results of our comparative study, we conclude:

(1) When the movement characteristics of objects are highly inconsistent over time, the variability in the segmentation results is high, resulting in a large number of short segments. Therefore, exploiting the MPC variability-based similarity (cf. Section 4.1) from the descriptive statistics of the segments works best for clustering trajectories with heterogeneous movement characteristics, that is, when objects exhibit high variability in the variation of their movement parameters over time (courier data, in our case).

(2) In contrast, the MPC sequence-based similarity (cf. Section 4.2) is better suited for clustering movement data with pronounced sequential movement behavior (as the hurricanes, in our case). The reason is that here the adapted string matching technique helps to detect similar sequences among segmented profiles.

Furthermore, the experiments suggest that applying the NWED similarity measure together with the MP class sequence representation of trajectories is better suited to study spatio-temporal behavior of moving objects in comparison to related spatial similarity measures (i.e. the EDM measure proposed by Chen *et al.* 2004). That is, as showed in the first case study, our approach is more effective in extracting the similarities in movement data w.r.t the evolution patterns of the objects' movement parameters over time (e.g. the speed behavior of hurricanes) in comparison to the method by Chen *et al.* (2004). Although in the experiments by using hurricane data at a regular sampling rate, we implicitly involved time in the computation of EDM to make it more comparable to NWED, the study showed that NWED is better suited to capture the spatio-temporal aspects of hurricane evolution. On the other hand, applying the geometric parameters such as turning angle, NWED provides comparable results to EDM, both being geometric

similarity measures.

Representing trajectories with a symbolic representation such as movement parameter class sequences (our approach), or movement pattern strings (MPS, Chen *et al.* 2004) significantly reduces the storage costs of trajectory data (e.g. to 12.5 % as shown in Chen *et al.* 2004). Moreover, the sequence representation is invariant to rotation and spatial transformations since the proposed segmentation algorithm relies on relative movement parameters computed between consecutive fixes along trajectories. The advantage of our proposed representation over MPS is that the number of classes and hence the domain of the sequences is much smaller (i.e. 9 classes in our case, 64 classes in MPS). As shown in Du Mouza *et al.* (2006), the pattern matching and retrieval costs are less for strings that are represented with a small number of characters (i.e. class labels in our case), especially in very large datasets. Moreover, similar to other edit-distance based approaches, our approach can deal with trajectories of unequal length and unequal sampling rate as well.

On the other hand, the string matching process is relatively slow and depends on the length of the profiles (i.e. $O(n^2)$). Therefore, the proposed methodology is computationally expensive for very large trajectory datasets with long trajectories. This issue is common to all edit distance-based approaches (e.g. NWED and EDM). In order to reduce the computational cost of similarity computation, Chen *et al.* (2004) proposed a *Modified Frequency Distance (MFD)* for frequency vectors that are obtained from movement pattern strings. This method is similar to our variability-based similarity approach (cf. Section 4.1). However, in the variability-based approach we employ more features in addition to the frequency of classes to describe characteristics of movement parameters. Besides, similar to variability-based clustering, MFD does not preserve the sequence of a movement. Another strategy to overcome this problem is to apply pruning approaches prior to the similarity computation (Chen *et al.* 2005).

Our proposed approach is originally developed for movement parameter profiles. However, the presented methodology can be applied for similarity analysis and clustering of other types of time series (since MP profiles can be seen as a specific type of time series). A similar method has been proposed for addressing threshold queries as well as similarity analysis in time series databases by Aßfalg *et al.* (2008). However, their method only considers the deviation (amplitude) of the time series. In contrast, our approach can handle the frequency of variations by considering the sinuosity of time series.

Finally, our study suggests there can hardly be a universally applicable similarity measure for movement trajectories. Depending on the application at hand, the best suited similarity measure and the adequate movement parameter must be chosen. Background knowledge about the investigated movement process helps making an informed choice. Such knowledge can come in the form of knowledge about the geographic context embedding the movement (as in the case of the traffic light in Experiment #3).

7. Conclusions and Future Work

In this paper, we introduce a new methodology for trajectory similarity detection, which moves similarity analysis beyond considering merely the geometric similarity of trajectories, towards considering movement dynamics. The method bases on the segmentation of movement parameter profiles of an object over time, which can be derived from trajectory data or directly observed using data from tracking sensors.

In our paper, we present a comparative evaluation, to show the usefulness of our approach in clustering both hurricane and courier trajectories. Besides, we experimentally

evaluate our approach in comparison to a relevant method by Chen *et al.* (2004). The experiments show that the proposed NWED similarity measure together with the MP class sequence representation of trajectories can be successfully applied in movement behavior analysis for finding structure in movement datasets. Particularly, when objects exhibit a common movement pattern, our approach becomes more effective in comparison to the available spatial similarity measures. We demonstrate that taking into account the domain and frequency variation of movement parameters can help identifying interesting patterns in the movement of objects.

Contrary to most existing work, the proposed similarity assessment approach focuses on the parameters describing the dynamic characteristics of movement, and does not deal with any geo-spatial or geometry-based similarity. As part of our future work, we plan to develop a combined approach that will integrate the proposed methodology with spatial/geometrical similarity analysis. Depending on the requirements of a particular application, one technique could be used as a filtering/pruning stage for the other. For example, in the second case study, we pruned courier subtrajectories to those along a selected route prior to the segmentation. Furthermore, this study shows that segmentation is a useful technique in extracting the structure of trajectories and assist knowledge discovery in movement data. However, our segmentation applies only one movement parameter at a time. As a future extension of our approach, we intend to develop a trajectory segmentation technique using multiple movement parameters. Moreover, we also intend to enrich the developed approach by incorporating contextual data. Finally, to alleviate the computational complexity of the proposed method, we are currently developing a multi-scale pruning procedure, working from coarse to finer spatial and temporal granularities.

8. Acknowledgment

This research is partly funded by the Research Fund ("Forschungskredit") of the University of Zurich. We would like to express our gratitude to Dr. Goce Trajcevski from the Department of Electrical Engineering and Computer Science, Northwestern University, for his valuable input at the initial stage of this research, and Jay Bregman (eCourier company, UK) for providing us with the courier data.

References

Alt, H., 2009. The Computational Geometry of Comparing Shapes. *In*: S. Albers, H. Alt and S. Näher, eds. *Efficient Algorithms.*, Vol. 5760 of *Lecture Notes in Computer Science* Springer Berlin, Heidelberg, 235–248.

Anagnostopoulos, A., *et al.*, 2006. Global distance-based segmentation of trajectories. *In*: *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining* ACM, 34–43.

Aßfalg, J., *et al.*, 2008. Similarity Search in Multimedia Time Series Data Using Amplitude-Level Features. *In*: *Advances in Multimedia Modeling*, 123–133.

Bozkaya, T., Yazdani, N., and Özsoyolu, M., 1997. Matching and indexing sequences of different lengths. *In*: *Proceedings of the sixth international conference on Information and knowledge management - CIKM '97* New York, New York, USA: ACM Press, 128–135.

Buchin, K., *et al.*, 2009. Finding long and similar parts of trajectories. *In*: *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - GIS '09* New York, New York, USA: ACM Press, 296–305.

Buchin, M., *et al.*, 2010. An algorithmic framework for segmenting trajectories based on spatio-temporal criteria. *In*: *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '10 New York, NY, USA: ACM, 202–211.

Chen, L., Özsu, M.T., and Oria, V., 2004. Symbolic Representation and Retrieval of Moving Object Trajectories. *In*: *MIR '04: Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval* In Proc. of the ACM SIGMM international workshop on multimedia information retrieval, 227–234.

Chen, L., Özsu, M.T., and Oria, V., 2005. Robust and fast similarity search for moving object trajectories. *In*: *SIGMOD '05: Proceedings of the 2005 ACM SIGMOD international conference on Management of data* New York, NY, USA: ACM, 491–502.

Ding, H., Trajcevski, G., and Scheuermann, P., 2008a. Efficient Similarity Join of Large Sets of Moving Object Trajectories. *In*: *15th International Symposium on Temporal Representation and Reasoning* Ieee, 79–87.

Ding, H., *et al.*, 2008b. Querying and mining of time series data: experimental comparison of representations and distance measures. *Proc. VLDB Endow.*, 1 (2), 1542–1552.

Dodge, S., Weibel, R., and Forootan, E., 2009. Revealing the physics of movement: Comparing the similarity of movement characteristics of different types of moving objects. *Computers, Environment and Urban Systems*, 33 (6), 419–434.

Dodge, S., Weibel, R., and Lautenschütz, A.K., 2008. Towards a taxonomy of movement patterns. *Information Visualization*, 7 (3-4), 240–252.

Doherty, S., *et al.*, 2001. Moving beyond observed outcomes: Integrating Global Positioning Systems and interactive computer-based travel behaviour surveys. *Transportation Research E-Circular, C*, 26, 449–466.

Du Mouza, C., Rigaux, P., and Scholl, M., 2006. On-line Aggregation and Filtering of Pattern-based Queries. *In*: *Scientific and Statistical Database Management, 2006. 18th International Conference on* IEEE, 333–342.

Du Mouza, C., Rigaux, P., and Scholl, M., 2007. Parameterized pattern queries. *Data & Knowledge Engineering*, 63 (2), 433–456.

Elsner, J. and Kara, A., 1999. *Hurricanes of the North Atlantic: Climate and society.* New York Oxford: Oxford University Press.

Etienne, L., Devogele, T., and Bouju, A., 2010. Spatio-Temporal Trajectory Analysis of Mobile Objects Following the Same Itinerary. *In*: *Proceedings of the International Symposium on Spatial Data Handling (SDH)*, Hong Kong.

Faloutsos, C., *et al.*, 1997. A signature technique for similarity-based queries. *In*: *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No.97TB100171)* IEEE Comput. Soc, 2–20.

Foster, G., 1996. Time series analysis by projection. I. Statistical properties of Fourier analysis. *The Astronomical Journal*, 111, 541–554.

Frentzos, E., Gratsias, K., and Theodoridis, Y., 2007. Index-based Most Similar Trajectory Search. *In*: *ICDE 2007. IEEE 23rd International Conference on Data Engineering*, 816–825.

Giannotti, F. and Pedreschi, D., 2008. *Mobility, Data Mining and Privacy.* Berlin Heidelberg: Springer-Verlag.

Kisilevich, S., *et al.*, 2010. *In*: *Spatio-Temporal Clustering : a Survey.*, Italy.

Laube, P., *et al.*, 2007. Movement beyond the snapshot Dynamic analysis of geospatial

lifelines. *Computers, Environment and Urban Systems*, 31 (5), 481–501.

Levenshtein, V., 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics-Doklady*, 10, 707–710.

Lin, B. and Su, J., 2005. Shapes based trajectory queries for moving objects. *GIS '05: Proceedings of the 13th annual ACM international workshop on Geographic information systems*, 21–30.

Little, J.J. and Gu, Z., 2001. Video retrieval by spatial and temporal structure of trajectories. *In*: *Proceedings of SPIE, the International Society for Optical Engineering* Spie, 545–552.

Malkin, W. and Holzworth, G.C., 1954. Hurricane Edna, 1954. *Monthly Weather Review*, 82 (9), 267–279.

Miller, H.J. and Han, J., 2009. *Geographic Data Mining and Knowledge Discovery*. Second Taylor & Francis Group.

Nanni, M. and Pedreschi, D., 2006. Time-focused density-based clustering of trajectories of moving objects. *Intelligent Information Systems*, 27, 267–289.

Pelekis, N., *et al.*, 2007. Similarity Search in Trajectory Databases. *In*: *TIME'07: Proceedings of the 14th International Symposium on Temporal Representation and Reasoning*, Jun.. Washington, DC, USA: IEEE Computer Society, 129–140.

Sinha, G. and Mark, D.M., 2005. Measuring similarity between geospatial lifelines in studies of environmental health. *Journal of Geographical Systems*, 7 (1), 115–136.

Trajcevski, G., *et al.*, 2007. Dynamics-aware similarity of moving objects trajectories. *In*: *GIS '07: Proceedings of the 15th annual ACM international symposium on Advances in geographic information systems* New York, NY, USA: ACM, 1–8.

van Kreveld, M. and Luo, J., 2007. The definition and computation of trajectory and subtrajectory similarity. *In*: *GIS '07: Proceedings of the 15th annual ACM international symposium on Advances in geographic information systems* New York, NY, USA: ACM, 1–4.

Vlachos, M., Gunopulos, D., and Kollios, G., 2002a. Discovering similar multidimensional trajectories. *In*: *ICDE '02: Proceedings 18th International Conference on Data Engineering* IEEE Computer Societ, 673–684.

Vlachos, M., Gunopulos, D., and Das, G., 2004. Rotation invariant distance measures for trajectories. *In*: *KDD '04: Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining* New York, New York, USA: ACM Press, 707 – 712.

Vlachos, M., Gunopulos, D., and Kollios, G., 2002b. Robust similarity measures for mobile object trajectories. *In*: *DEXA'02: Proceedings of the 13th International Workshop on Database and Expert Systems Applications* Los Alamitos, CA, USA: IEEE Computer Society, 721 – 728.

Yan, Z., *et al.*, 2010. A Hybrid Model and Computing Platform for Spatio-Semantic Trajectories. *In*: *The Semantic Web: Research and Applications* Springer, 60–75.

Yanagisawa, Y., Akahani, J.i., and Satoh, T., 2003. Shape-Based Similarity Query for Trajectory of Mobile Objects. *In*: *MDM '03: Proceedings of the 4th International Conference on Mobile Data Management* London, UK: Springer-Verlag, 63–77.

Yujian, L. and Bo, L., 2007. A normalized levenshtein distance metric. *IEEE transactions on pattern analysis and machine*, 29 (6), 1091–1095.